3D Visual Learning for Real-World Scenarios

Candidate: Antonio Alliegro

Summary

Thanks to the advancement of deep learning over the past few years and the recent hype on natural language and vision-language models, we are getting closer to artificial intelligent agents that can converse and interpret their surroundings. However, whether they can seamlessly navigate and interact within the real-world remains an open question. Indeed, these capabilities require a nuanced understanding of complex 3D environments which remain challenging for autonomous systems. Distributional shifts are among the hurdles that still need to be overcome: they occur when an autonomous system is deployed in environments significantly different from their training data, resulting in reduced performance and limited applicability in safety-critical scenarios. These shifts can include changes in weather, lighting, and style, as well as the emergence of novel semantic concepts.

This thesis tackles the pivotal issue of distributional shifts in 3D computer vision, with a focus on the domain and semantic variations that autonomous systems encounter in the real-world, where the variability and complexity far exceed those of controlled, synthetic training environments. It proposes a collection of innovative methodologies and studies aimed at enhancing the adaptability and resilience of autonomous systems to these shifts.

The first part (Chapters 3, 4, 5) is dedicated to the development of robust 3D vision models able to tackle the Synthetic to Real domain shift. In the studied scenarios, 3D point clouds are captured by on-site sensors and differ substantially from synthetic training data. They face challenges such as occlusions, interactions with other objects, sensor noise, and background clutter. Chapter 3 introduces a novel framework for 3D Domain Adaptation and Generalization through Self-Supervision. It showcases the benefits of combining supervised and self-supervised learning signals, improving feature representations for various 3D vision tasks, addressing the synth-to-real *domain shift* and *scarcity of annotated data*. To further address the Synthetic to Real domain gap, we design a deep learning framework that is able to *complete partial point clouds*. We name it DeCo and present its details in Chapter 4. It leverages denoising and contrastive learning self-supervised pretexts to achieve *category-agnostic completion* and *robustness to different types of input corruption*. In Chapter 5, we focus on object manipulation from real-world 3D point clouds.

In particular, we develop an end-to-end learning strategy, Learning to Grasp (L2G), for generating 6-DOF grasps from partial object point clouds, showcasing *robustness to real-world 3D scans* and *generalization to novel object categories*.

The second part of this thesis focuses specifically on the challenges of Semantic Shift faced by 3D vision algorithms deployed in the real-world. This type of distributional shift arises when systems come across new object categories and semantic concepts that were not part of their training data. In particular, Chapter 6 introduces 3DOS, the *first in-depth investigation of 3D Open Set Recognition and Semantic Novelty Detection.* This study evaluates how well 3D learning models can identify novel semantic concepts. We adapt and test a wide range of state-of-the-art methods from 2D Out-Of-Distribution detection and Open Set Recognition literature, examining their performance on 3D data in a series of progressively challenging testbeds. Furthermore, Chapter 7 presents 3D-SeND, *a novel method for 3D Semantic Novelty Detection that does not require a learning phase on task-specific support set data.* This approach leverages the powerful representational capabilities of large-scale pre-trained models, combined with reasoning on the compositionality of 3D shapes. It is specifically designed to be *efficient* and for *seamless application into diverse real-world scenarios*, as it removes the requirement for specialized learning stages tailored to the specific downstream task.

In conclusion, this thesis offers practical solutions for a wide range of real-world 3D tasks, including recognition, segmentation, completion, and manipulation of point clouds. Additionally, it addresses the need for autonomous systems to detect semantic novelty in open-world scenarios, proposing the first 3D Open Set Recognition benchmark and an efficient novelty detection approach that bypasses the requirement for task-specific training or fine-tuning. The insights and methodologies presented here are set to significantly impact the field of 3D visual learning and its practical applications across real-world scenarios.